Where to open a Spanish restaurant in Madrid

Project Report

SEPTEMBER 2019

PREPARED FOR
IBM Data Science Professional Certificate
Applied Data Science Capstone
Peer Graded Assignment: Capstone Project – The Battle of the Neighborhoods
https://github.com/mnlcsty/Coursera_Capstone

PREPARED BY
Manuel Costoya Ramos
manuel.costoya.ramos@gmail.com
https://www.linkedin.com/in/manuel-costoya-ramos

Table of Contents

Business Problem

Madrid is a big city. According to Wikipedia it is the third most populous city in the European Union behind Berlin and London. It has 131 boroughs grouped in 21 districts. It is a lively city where locals like to eat out, dine out, and in general go out. And it is now a popular travel destination for tourists from across the globe, so much so that in its September 2018 issue Time Out magazine named one of its neighborhoods “the coolest neighborhood in the world”.

Any time of the year, any day of the week, the Centro, Retiro and Salamanca districts are swarming with tourists and locals. The remaining districts inside the M-30 inner beltway (Arganzuela, Chamberí, Chamartín and Tetuán) are also very dynamic but the tourist presence is traditionally somewhat lower.

With so many locals and visitors eating out and dining out, it would seem that restaurants must be full all the time. And judging from the table occupancy of pavement cafés, that would seem to be the case. But the tables inside the restaurant frequently tell a very different story. More often than not they are all empty. Get a table in a pavement café if you are lucky enough, and you will realize that many people sit there for hours in front of a up of coffe, a beer or a bottle of water.

Complicating things further, a “restaurant bubble” has been growing in Madrid for some years, with more and more restaurants opening and closing every year. Popular gastronomy critic El Comidista dedicates this article to the problem.

As it happens the restaurant business is as hard in Madrid as anywhere else. A lot of people have lost money, gotten into debt, gone bankrupt and/or seriously damaged their family relationships after opening a restaurant attracted by the apparent business simplicity and movement of people, to find soon after how difficult the business really is.

The restaurant industry and its analysts have long studied the reasons why restaurants fail. Poor choice of location is among the most important, as discussed in the Literature Review section, and Data Science can help mitigate the associated risks.

This report is directed at anyone facing the choice of location for a new Spanish restaurant in Madrid.

Literature Review

Among the many articles and blog posts about the reasons for restaurant failure that can be found online, the one that stands out is “Why Restaurants Fail” by H.G. Parsa, John T. Self, David Njite and Tiffany King from Cornell University. In one of the sections the authors list the following elements of restaurant failure: lack of documented strategy, seat-of-the-pants management, management operations by “putting out fires”, focusing on only one aspect of the business, poor choice of location, restaurant concept and location mismatch, insufficient start-up or operational capital, lack of business experience or knowledge of restaurant operations, poor communication with customers, negative customer perception of value, inability to mantain operational standards, loss of authenticity (for ethnic restaurants), becoming everything to everyone, underestimating the competition, lack of owner commitment due to family demands, lack of operational performance evaluation systems, frequent changes in management, tardy establishment of vision and mission statement, failure to maintain management flexibility and innovation, noncontrollable external factors, and entrepreneurial incompetence. Restaurant density is also discussed as an ambivalent factor: it can be beneficial to a point as it attracts more traffic, but it is strongly correlated with restaurant turnover.

The article conclusions seem as valid today as they were back in 2005 when it was written with one significant addition: social networking as a distinct and all-important form of customer relationship management.

Broadly speaking, the reasons for restaurant failure fall into one of the following categories:

- Lack of owner commitment, business knowledge and/or experience
- Insufficient capital
- Poor management, customer relationship marketing and/or social networking
- Loss of concept
- Poor choice of location

Of the above categories for restaurant failure , the one Data Science can help with is the choice of location. A model can be built to cluster all posible locations based on restaurant density, placement-related capital requirements, and direct competition—or suitable proxies. The best cluster(s) can then be chosen based on the effect of restaurant density and competition on failure risk and the owner’s concept, decision criteria and access to capital. Depending on how the model features are distributed, one or several equally attractive locations may stand out as the best choices within the chosen clusters, or additional input by the owner may be required in order to make a choice.

Data Sources

Madrid neighborhood list is available on the following Wikipedia page:
https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid.
The page was scraped to extract for each neighborhood: district, neighborhood number, neighborhood, area and link to the neighborhood Wikipedia page.

The Wikipedia page for each neighborhood was in turn be scraped to extract the neighborhood latitude and longitude. For example, the page corresponding to Embajadores neighborhood is: https://es.wikipedia.org/wiki/Embajadores_(Madrid).

The FourSquare 'explore' endpoint was be used to extract the most popular nearby venues for each neighborhood. The data was be used to compute neigborhood popularity and direct competitive pressure, both model features. The choice of the 'explore' over the 'search' endpoint is justified in the Methodology section.

Finally, the 2nd hand housing prices for each neighborhood were obtained from the following file published by the City Council:
https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Edificaci%C3%B3n%20y%20Vivienda/Mercado%20de%20la%20Vivienda/Precios%20de%20la%20Vivienda/Distritos/E3320219.xls.
The choice of this price index as a model feature is justified in the Methodology section.

Methodology

As discussed in the Literature Review section, a model can be built to cluster all posible locations based on restaurant density, placement-related capital requirements, and direct competition—or suitable proxies.

This section discusses the variables and methods chosen to build that method, and how they helped solve the problem at hand.

Variables

Restaurant density and neighborhood popularity

In the Literature Review restaurant density was identified as an ambivalent factor: it can be beneficial to a point as it attracts more traffic to a location, but it is strongly correlated with restaurant turnover.

For this project there was no readily available public data on restaurant density for each neighborhood. Two alternative sources were then considered: the Foursquare API 'search' and 'expore' endpoints.

The FourSquare 'search' endpoint allows a targeted search of all venues in category "Food" and all its sub-categories (categoryId= 4d4b7105d754a06374d81259) within a radius of a given location. But it returns up to 50 results per query, and the API documentation does not confirm whether these are the 50 most popular or any random results.

On the other hand, the FourSquare 'explore" endpoint returns up to 100 results per query[1], the venues returned are the most popular within a radius of the given location.

More popular venues in a location attract more traffic. So "popular venue density" or **neighborhood popularity** is a good proxy for restaurant density, and was used as one of the model features.

It was computed as the total count of popular venues returned by the FourSquare 'explore' endpoint within a 500m radius of each neighborhood coordinates.

A neighborhood popularity histogram was plotted to confirm sufficient variability accross neighborhoods.

Location-related capital requirements and 2nd hand house pricing

In the Literature Review section, insufficient start-up or operational capital was identified as one of the elements of restaurant failure.

For the purposes of this project, only location-related capital requirements were relevant. Unfortunately, no public data was found for rental or second hand commercial property price averages for Madrid neighborhoods.

But **2nd hand housing price** is an almost *de facto* standard to assess the cost of land in Spain. And while not directly valid as an estimate of location-related capital requirements, it is no doubt strongly correlated with them. Thus it can be used as a proxy for the purposes of this project and was included as a model feature.

A 2nd hand housing price histogram was plotted to confirm sufficient variability accross neighborhoods.

Direct competition

In the Literature Review section, loss/lack of a clear concept was identifed as one of the elements of restaurant failure.

For this project the restaurant owner's concept was assumed to be a Spanish restaurant. Consequently, the following FourSquare venue categories were considered direct competitors:

1 The following tests were done to confirm this:
- No limit in the url: 30 results returned, issues with the response
- Limit > 100: 100 results returned

- “Spanish Restaurant” and its sub-categories “Tapas Restaurant” and “Paella Restaurant”
- “Restaurant”, as most generic restaurants in Madrid include at least some spanish cuisine on their menu

Direct competition was computed as the total sum of the relative frequencies of these venue categories in each neighborhood, and was included as a model feature.

The data source for features ‘direct competition’ and ‘neighborhood popularity’ was the same: he popular venues returned by the FourSquare ‘explore’ endpoint within a 500m radius of each neighborhood coordinates

A direct competition histogram was plotted to confirm sufficient variability accross neighborhoods.

Methods

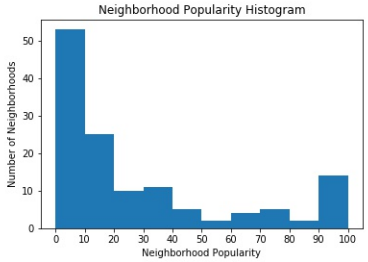
The methods chosen to build a clustering model of the neigborhoods where:

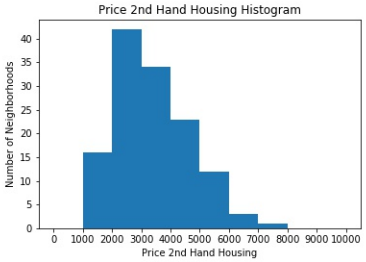
- *Histogram* plots to confirm sufficient feature variability accross neighborhoods
- *Scikit-Learn’s MinMaxScaler* to normalize the features
- *Scikit-Learn’s KMeans* to cluster the neighborhoods
- Feature *quartiles, means* and *standard deviations* to analyze and describe the resulting clusters in a meaningful way for the purposes of the project

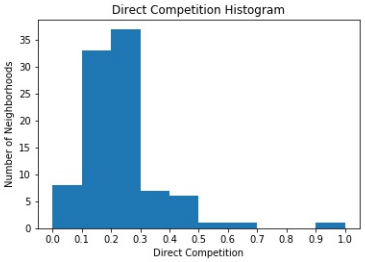
The cluster description table was used in combination with the owner’s preferences to solve the problem at hand, as described in the Discussion section.

Results

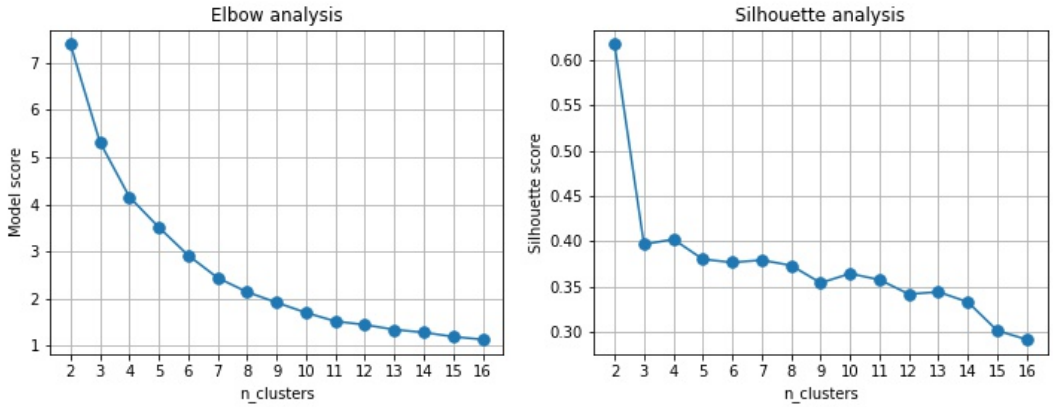
The following histograms allowed to visually confirm sufficient variability accross neighborhoods for all three features:



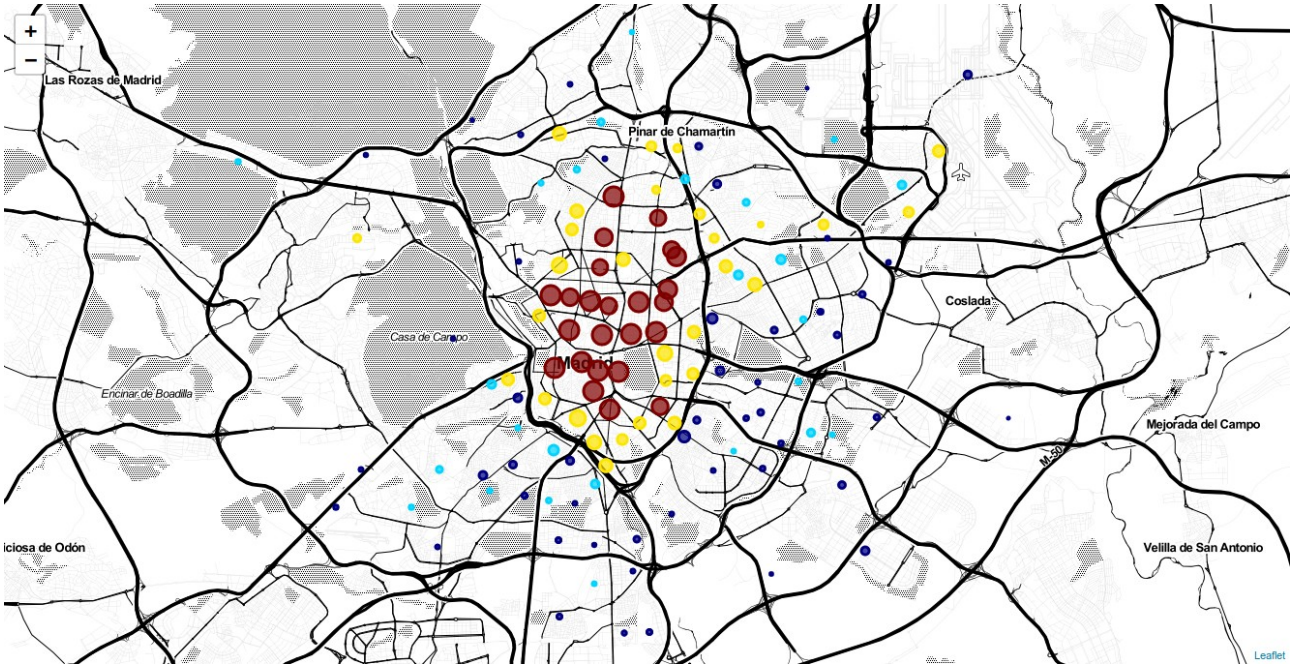




To tune the *n_clusters* parameter of *Kmeans,* two approaches were combined: the elbow and silhouette analyses. Both suggested that 4 was the optimal parameter value for *n_clusters:*



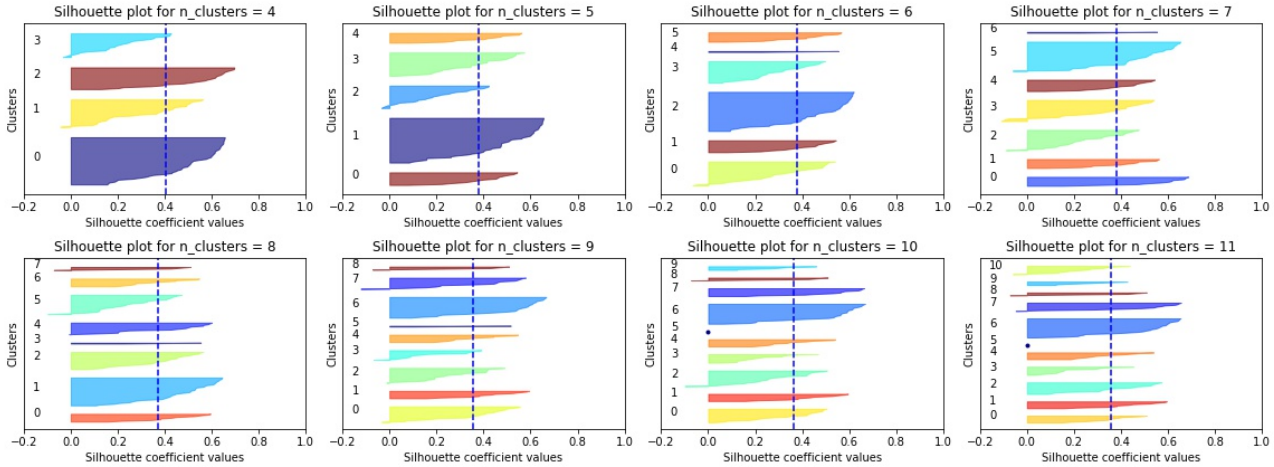
A KMeans clusterer was intialized with *n_clusters*=4 and fit to the model features. To better visualize the results, a map with all neighborhoods was plotted using the *Folium* package. Neighborhoods were plotted as circle markers with an area proportional to neighborhood popularity. A different color was used for each cluster—jet colormap with blue to red for less popular to more popular:



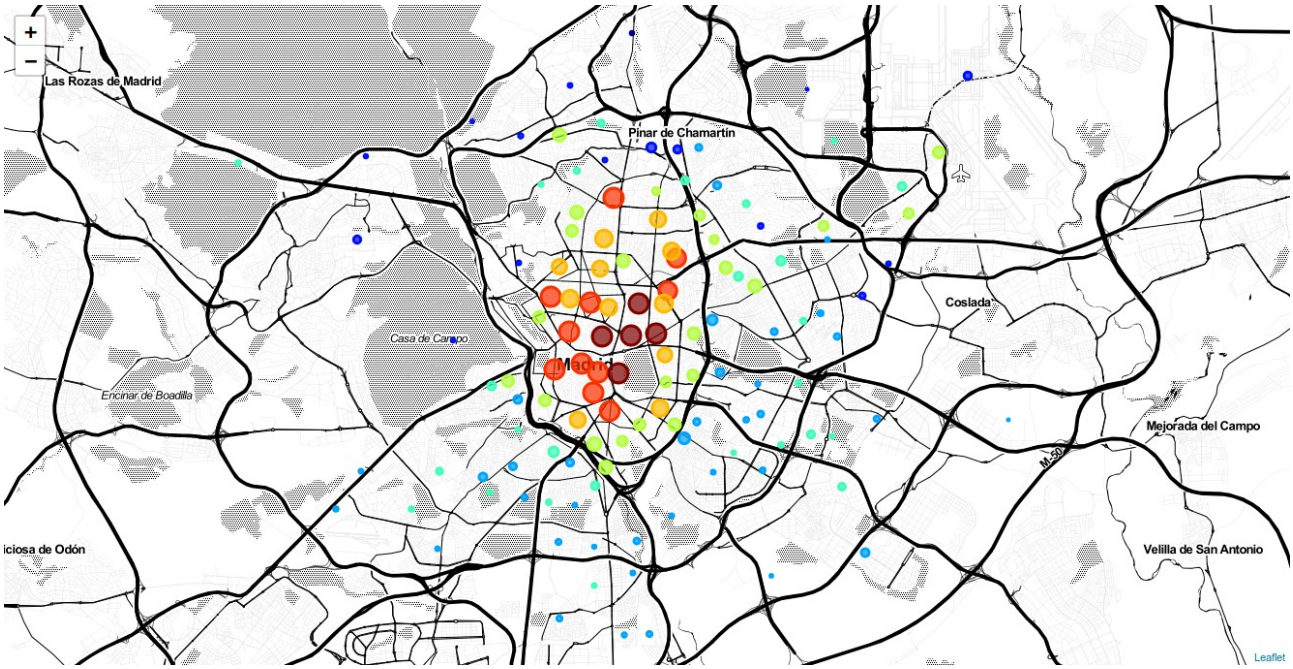
The map showed that all popular neighborhoods (larger markers) were included in the same cluster (deep red). This was considered too coarse for the purposes of the project.

A larger value of *n_clusters* was required to achieve a finer clustering, but neither the elbow nor the

silhouette score were of help for values of *n_clusters* above 4. To pick a higher value, the approach chosen was misclassification risk minimization. Silhouette plots were produced for values of *n_clusters* between 4 and 11 with the same color code—jet colormap with blue to red for less popular to more popular:



The plots suggested that *n_clusters*=8 was the simplest, cleanest model to split the popular neighborhoodsaccross at least three clusters (short silhouettes for dark red, red and orange clusters). Thus a KMeans clusterer was initialized this time with *n_clusters*=8 and fit to the model features, and the following map was plotted to better visualize the new clusters:



The new map confirmed that the popular neighborhoods (larger markers) were split accross three different clusters (deep red, red and orange on the map) . This level of detail was considered adequate for the purposes of the project.

Finally, the clusters were described by comparing cluster feature averages with total feature quartiles, means and standard deviations to produce the following **cluster description table**:

Cluster	Number of neighborhoods	Popularity	Price 2nd Hand Housing	Direct Competition
7	5	Very high	Very high	Moderate
0	11	Very high	High	Moderate
6	11	High	High	High
2	23	Moderate	Moderate	Moderate
5	26	Below average	Below average	High
1	37	Below average	Low	Low
4	16	Below average	Moderate	Low
3	2	Low	Average	Very high

Discussion

This section describes how the analytical results were used in combination with the restaurant owner's preferences to solve the problem at hand.

The restaurant owner's preferences were assumed to be:

- Neighborhood popularity: high popularity is desirable, very high popularity or less than high popularity is not
- Location-related capital requirements: avoid prime commercial property
- Direct competition: the lower the better in general, but check for causes of unsuitability if very low (example: a popular neighborhood with no direct competition where all restaurants are ethnic)

To identify the most suitable cluster(s), the owner's preferences where simply applied to the cluster description table. Cluster 6 was identified as the most suitable:

Cluster	# Nbhs.	Popularity	Price 2nd Hand Housing	Direct Competition	Recommendation on the basis of owner's preferences
7	5	Very high	Very high	Moderate	Too popular, too expensive
0	11	Very high	High	Moderate	Too popular, expensive
6	**11**	**High**	**High**	**High**	**RECOMMENDED**
2	23	Moderate	Moderate	Moderate	Not popular enough
5	26	Below average	Below average	High	Not popular enough
1	37	Below average	Low	Low	Not popular enough
4	16	Below average	Moderate	Low	Not popular enough
3	2	Low	Average	Very high	Not popular enough, too much direct competition

To check for possible causes of unsuitability, the most popular venue categories for the neighborhoods within Cluster 6 were ranked and arranged in the following table:

Rank	1.0	2.0	3.0	4.0	5.0
Neighborhood					
Acacias	Spanish Restaurant	Park	Bar	Pizza Place	Café
Almagro	Restaurant	Spanish Restaurant	Italian Restaurant	Bar	Plaza
Arapiles	Spanish Restaurant	Bar	Bakery	Café	Restaurant
Ciudad Jardín	Tapas Restaurant	Bakery	Bar	Café	Coffee Shop
Cuatro Caminos	Spanish Restaurant	Italian Restaurant	Bakery	Tapas Restaurant	Burger Joint
Hispanoamérica	Spanish Restaurant	Bar	Restaurant	Grocery Store	Pizza Place
Ibiza	Spanish Restaurant	Restaurant	Tapas Restaurant	Italian Restaurant	Seafood Restaurant
Lista	Spanish Restaurant	Restaurant	Seafood Restaurant	Coffee Shop	Bar
Pacífico	Spanish Restaurant	Bar	Café	Grocery Store	Pizza Place
Ríos Rosas	Tapas Restaurant	Café	Italian Restaurant	Japanese Restaurant	Restaurant
Vallehermoso	Spanish Restaurant	Bar	Restaurant	Bakery	Café

No causes for unsuitability were found in any of the neighborhoods.

To pick the best neighborhood(s) within the cluster to start the search for premises, feature averages were computed for the neighborhoods within the cluster, then the neighborhoods were ranked by average popularity and arranged in the following table for easier comparison:

		Popularity	Price 2nd Hand Housing	Direct Competition
Cluster	**Neighborhood**			
6	**Lista**	78	5498.0	0.307692
	Arapiles	76	4855.0	0.223684
	Cuatro Caminos	75	4115.0	0.200000
	Ciudad Jardín	72	4321.0	0.152778
	Almagro	70	6175.0	0.271429
	Pacífico	69	4059.0	0.217391
	Hispanoamérica	64	5024.0	0.312500
	Acacias	62	4046.0	0.225806
	Ríos Rosas	62	5044.0	0.209677
	Vallehermoso	57	4341.0	0.245614
	Ibiza	52	5226.0	0.442308

One neighborhood stood out among the others:

Neighborhood	Cluster	Popularity	Price 2nd Hand Housing	Direct Competition
Cuatro Caminos	6	High	Above average (1 level below cluster)	Above average (1 level below cluster)

For the presentation to the restaurant owner, the following map including only the neighborhoods in Cluster 6 was plotted:

And **Cuatro Caminos** was recommended as the neighborhood to start looking for premises for its high popularity, moderate second hand housing prices and also moderate direct competition.

The most significant caveat identified for the chosen approach was the dependence of neighborhood popularity on the quality and representativity of FourSquare data, and the limitations of its API 'explore' endpoint. Alternatives like Google could be more representative or up-to-date in some places, or provide unlimited access to endpoint query results.

Conclusion

The restaurant business is hard and there are many elements of restaurant failure. As shown in this report, Data Science can help with one of them: poor choice of location.

The specific case of a new Spanish restaurant in Madrid was analyzed for this project, a cluster of 11 neighborhoods was identified as the most suitable area given the owner's preferences, and one neighborhood within the cluster was chosen as the starting point for the search of premises.

The same principles can be applied to any other restaurant category in any other city provided that the the effect of location popularity on traffic and restaurant turnover is understood, categories in direct competition can be identified, and data sources are available for the model variables or suitable proxies.